Learned Student Models with Item to Item Knowledge Structures

Michel C. Desmarais (michel.desmarais@polymtl.ca), Peyman Meshkinfam (peyman.meshkinfam@polymtl.ca) and Michel Gagnon (michel.gagnon@polymtl.ca)*
École Polytechnique de Montréal

Abstract. Probabilistic and learned approaches to student modeling are attractive because of the uncertainty surrounding the student skills assessment and because of the need to automatize the process. Item to item structures readily lend themselves to probabilistic and fully learned models because they are solely composed of observable nodes, like answers to test questions. Their structure is also well grounded in the cognitive theory of *knowledge spaces*. We study the effectiveness of two Bayesian frameworks to learn item to item structures and to use the induced structures to predict item outcome from a subset of evidence. One approach, POKS, relies on a naive Bayes framework whereas the other is based on the Bayesian network learning and inference framework. Both approaches are assessed over their predictive ability and their computational efficiency in different experimental simulations. The results from simulations over three data sets show that they both can effectively perform accurate predictions, but POKS generally displays higher predictive power than the Bayesian network. Moreover, the simplicity of POKS translates to a time efficiency of one to three orders of magnitude greater than the Bayesian network runs. We furhter explore the use of the item to item approach for handling concepts mastery assessment. The approach investigated consist in augmenting an initial set of observations, based on inferences with the item to item structure, and feed the augmented set to a Bayesian network containing a number of concepts. The results show that augmented set can effectively improve predictive power of a Bayesian network for item outcome, but that improvement does not transfer to the concept assessment in this particular experiment. We discuss different explanations for the results and outline future research avenues.

Keywords: Student models, Probabilistic models, Bayesian networks, Bayesian inference, POKS, Knowledge spaces, Knowledge assessment, Adaptive testing, CAT, Empirical simulations

1. Introduction

Student models are at the core of intelligent and adaptive learning environments. We expect from these models that they provide an accurate and often fine grained assessment of the user's concepts mastered and, sometimes, even misconceptions. Graphical probabilistic models such as Bayesian Networks (BN) are often used because they have many of

* This work has been supported by the National Research Council of Canada.

the qualities we look for in student models. Each node in a graphical model structure can represent a specific concept, skill, or misconception, thereby allowing detailed knowledge assessment. Each node can gather evidence from complex relationships between other nodes to yield a probabilistic assessment of mastery. They can be learned from data or engineered by domain experts when data is not available.

However, graphical student models are seldom completely learned from data because they contain numerous concepts that cannot be observed directly. Learning models with a significant number of hidden nodes is challenging. Most of the research work reported in the student modeling literature make use of human expertise at some point in the modeling process to link concept between themselves and to observable items.

That reliance on human expertise to model domain knowledge and perform knowledge assessment can be a major drawback in many contexts. It requires domain experts, and sometimes probabilistic modeling experts too, each of which can be a bottleneck. The reliance on the joint contribution of domain and cognitive/probabilistic modeling experts is doomed to be a major hindrance in a practical context.

In light of these practical constraints, the need for designing student models that can be learned from data is very high in our view. Some techniques such as IRT (Item Response Theory) and Multidimentional IRT (Wang and Chen, 2004) do offer means to automate at least some part of the effort and relieve its dependency on modeling expertise, but they cannot handle many concepts, nor links between concepts themselves. They lack the highly flexible modeling offered by graphical models, such as a BN. Many applications that require detailed student models tend to rely on BN because of this flexibility (see for eg. VanLehn et al., 2005).

We will review the issues that need to be taken into account for building student models that can achieve flexible, fine grained assessment, while relying as much as possible on an automated learning approach.

Later, we discuss the different avenues for building student models from data and, in particular, the approaches that rely on item to item probabilistic structures. Let us first introduce the issues that surround the choice of a student modeling approaches.

2. Tradeoffs Between Student Models

As the success of intelligent learning environments grow, and as some of them become used in real-life settings (see for eg., VanLehn et al.,

Mitrovic et al., 2005, 2001), researchers in the field have come to focus attention on the need to build rapid, cost-effective means of developing these environments (Aleven et al., 2006; Martin and Mitrovic, 2001; Kodaganallur et al., 2005). This is currently a major issue. Aleven et al. (2006) report that authoring time for one hour of instruction of an intelligent tutoring ranges from 200–300 hours. We thus need to reduce this cost and better understand the tradeoffs we face when choosing one paradigm over another.

The need for more effective means of building these environments explains, in part, the accrued interest for machine learning and data mining techniques to support their development. Learned student models are part of this trend. We review some of the issues and qualities between learned and knowledge engineered approaches and between different learned models themselves.

- **Flexibility and expressiveness** As emphasized above, AI-based learning environments often rely on fine-grained assessment of abilities and misconceptions. Graphical probabilistic models are highly suitable for fine-grained cognitive diagnostic. This quality may explain why they enjoy a higher visibility in the student modeling field than does the classic psychometric approach, namely IRT, which is tailored to classifying an examinee as *master* or *non-master* (sometimes *undecided* too) and over a few dimensions.

  Beyond graphical models, other approaches based on using constraints to learn domain rules have also been investigated with encouraging results (Suraweera et al., 2005; Mitrovic et al., 2001). They can yield detailed diagnostics while relying on rules induced from data. Although we focus our discussion on graph models in the current paper, the reader should note that the issues raised here could apply to such models as well.

- **Cost of model definition** Models such as those found in BNs (see, for example Vomlel, Conati et al., 2004, 2002) can require considerable expert modeling effort. This modeling effort can be well worthwhile in academic domains such as mathematics or physics, but it can prove overly costly for many other applications. On the contrary, data driven approaches can completely waive the knowledge engineering effort when data is available in sufficient quantity for the learning approach chosen.

- **Scalability** The number of concepts/skills and test items that can be modeled in a single system is another factor that weights into evaluating the appropriateness of an approach. Taking IRT as an example again, it allows good scalability to very large tests

(hundreds of items). However, IRT does not attempt to link the hidden, latent skills among themselves like many graph models do. Techniques to manage the complexity are thus required. The hierarchical structure of concepts of many graph models is an efficient means towards this goal, but it has to be defined by an expert and cannot be learned from data, at least currently.

– **Cost of updating** The business of skills assessment is often confronted with frequent updating to avoid over exposure of the same test items. Moreover, in domains where the skills evolve rapidly, such as in technical training, new items and concepts must be introduced regularly. Approaches that reduce the cost of updating the models are at significant advantage here. This issue is closely tied to the knowledge engineering effort required and the ability of the model to be constructed and parametrized with a small data sample.

– **Accuracy and reliability of prediction** Student modeling applications such as Computer Adaptive Testing (CAT) are critically dependent on the ability of the model to provide an accurate assessment with the least number of questions. Models that can yield confidence intervals, or the degree of uncertainty of their inferences/assessment, are thus very important in this field as well as in many context in which measures of accuracy is relevant. The requirement over accuracy and confidence measures is well argued by Horvitz (1999) for the case of intelligent user agents and it also applies for learning agents: there often is a high cost associated with a wrong system intervention/decision based on an incorrect diagnostic. The user looses trust in the system's recommendations, or whatever "intelligent feature" it displays, and can get annoyed to the point of loosing interest. We therefore need to assess the value of a system's decision in light of the confidence we have in the diagnostic, and refrain from taking action unless we have a sufficient level confidence.

– **Reliability and sensitivity to external factors** A factor that is often difficult to assess and overlooked is the reliability of a model to environmental factors such as the skills of the knowledge engineer, the robustness to noise in the model, and to noise in the data used to calibrate a model.

  Handcrafted models, in particular, are subject to idiosyncrasies and human biases. They cannot readily offer means to predict their reliability. Whereas extensive research in IRT has been conducted

to investigate reliability and robustness under different conditions, little has been done in intelligent learning environments yet.

While knowledge engineering approaches can suffer from the expert's biases, learned approaches can also suffer from over-fitting to a particular sample's idiosyncratic characteristics, such as the specific teaching from a given teacher, a given school, or for a given year. Sampling issues must remain a concern for all learned approaches.

– **Mathematical and theoretical foundations** The advantages of formal and mathematical models need not be defended. Models that rely on sound and rigorous mathematical foundations are generally considered better candidates over *ad hoc* models because they provide better support to assess accuracy and reliability, and they can often be automated using standard numerical modeling techniques and software packages.

  The same argument can be made for the cognitive foundations of a student model. For example, a sound theoretical foundation in psychometric measurement has been developed behind IRT. Similarly, the theory of *knowledge spaces* (Doignon and Falmagne, 1999) also offers a strong formal basis for student model upon which a few models are based, including the POKS framework we will discuss later (section 3).

– **Approximations, assumptions, and hypothesis** In the complex field of cognitive and skill modeling, all models must make a number of simplifying assumptions, hypothesis, or approximations in order to be applicable. This issue is closely linked to reliability and sensitivity. Some approach may work well in one context and poorly in another because of violated assumptions.

– **Re-usability** The re-usability of a student model can represent a critical quality to strive for. Reusable student models share the same qualities as reusable code: the initial cost is amortized over the number of times it is reused, and testing and validation needs to be done only once. The work of Zapata-Rivera and Greer (2004) is an example of an architecture that lends itself towards this goal of reusing student models and where the underlying student models are Bayesian. Generic user models also share the same goal (Kay et al., 2002; Kobsa, 2001). Moreover, the re-usability factor can outweigh the disadvantages of a knowledge engineered model if it can be reused to a sufficient extent. However, it appears we are still a few years away before any standard can emerge and that

could enable generic and reusable student models. Nevertheless, we do study a means of combining a learned model with an existing BN model in section 8 with a perspective that could lead to re-usability.

The above mentioned factors will determine the practical value of a student modeling approach. Ideally, we would like to reuse as much as possible existing student models, and rely over a fully automated model learning approach that requires little data to build and calibrate. Yet, we also want a model that yields detailed and accurate knowledge assessment. Such an approach would limit the effort of model building to that of data gathering. It would also facilitate model update. For example, adding new test items and new concepts would only require re-running the learning algorithm, which could in principle be done in real time as new data is gathered. Finally, given an algorithmic approach to model building, reliability and accuracy is less dependent on environmental factors, such as human subjectivity and expertise, and it can be assessed either through parametric methods or through non parametric methods such as the bootstrap approach.

Currently, and along with the constraint based approaches mentioned earlier (Suraweera et al., 2005; Mitrovic et al., 2001), the prevalent student modeling frameworks that can yield fine grained assessments are the graphical models and, in particular, the Bayesian Network framework (see, for example Vomlel, Conati et al., 2004, 2002). Graphical models score high on the flexibility factor and it has a strong mathematical foundation with plenty of tools and libraries available. Nodes of a graphical model can represent detailed cognitive diagnostics such as misconceptions. They can be linked together in a flexible manner that can express complex non-linear relations.

On the other hand, graphical models can suffer important shortcomings with regards to the other dimensions. The need for a knowledge engineering effort to build graphical models can significantly impact the cost of model definition, which in turn will influence the scalability and cost of updating. In the context of technical training, where skills evolve rapidly, this factor weighs a lot because the learning content and the assessment tests require frequent updates. Moreover, having a human intervention in the process of model building and calibration makes it more difficult to predict how the model will perform, as it becomes dependent on the modeler's skills. Knowledge engineered intensive models can only be considered in the context where re-usability or a sufficient amortizing period can justify the initial investment.

However, learned models can relieve the shortcomings of knowledge engineered models. In general, it is far easier to gather and use data

to build and calibrate a model than to rely on an expert in both the domain content and the probabilistic technique involved to go to the process of building such model. Moreover, parametric or non parametric statistics can be developed to predict the accuracy and reliability of the model. Of course, the amount of data to build a learned model is critical and, for each context of application, there comes a point where the data gathering effort can outweigh the benefits. Learned models that require small data sets are obviously at an advantage in most contexts.

That said, care must be taken to ensure the sampling is representative. Possibly the highest risk of learned models is that of sampling problems which can lead to invalid results that go unnoticed. For example, a specific teacher (or a school) can select and order the teaching material differently from another teacher (or a school from some another region or country). These peculiarities will be tapped by the learned models and incorrectly used for inference in a different context.

Notwithstanding the sampling pitfall of learned models to student modeling, their advantages are compelling and the search for such models should be paramount in our view. However, a universal model that meets all the above requirements is an evasive goal because there are tradeoffs between some of these requirements. For example if a very large training data set is available, we can model complex interrelationships by straightforward conditional probability tables that make no simplifying assumptions. With small data sets, we often have to make different assumptions to counter the lack of data, or revert to subjective estimation of parameters. The validity of the assumptions and the sensitivity of the results to estimation errors will vary on a case by case basis. When making assumptions and approximations, it becomes important to understand how they can affect the results. A good understanding of the circumstances under which a given model is best suited in a given context is most likely a more appropriate goal than that of looking for a universal model. The above dimensions could serve as a basis for investigating this issue.

This paper proposes an approach that addresses many of the issues outlined. It builds upon the fact that observable items have a structure among themselves that can readily be induced from data, and further shows how to use this structure as a first step in a knowledge assessment process.

In the next section, we review two models that build item to item structures and that are learned from small data samples. One model is based on BN structural learning techniques whereas the other is based on the POKS framework that relies on a local independence assumption. Both frameworks offer detailed knowledge assessment by estimating the mastery of individual *knowledge items*, but they do not

incorporate concepts that can include many knowledge items and sub-concepts. We further discuss how concepts can be introduced within an item to item framework by reusing an existing BN for that purpose.

3. Item to Item Node Structures and the Theory of Knowledge Spaces

Student models are generally organized as a hierarchy of concepts with observable nodes, namely test items, as leaves of this hierarchy. The "non observable" nodes are concepts, skills, and misconceptions. They are considered hidden nodes in the sense that they cannot be directly observed. The hierarchy can also include relations that break away from a pure hierarchy and link sibling nodes together, or link children nodes to multiple parents. Nevertheless, the general underlying organization remains hierarchical. Even IRT and MIRT can be conceived as hierarchical graph models consisting of one skill node (or more for MIRT) linked to any number of test items as children.

One family of models departs from the hierarchical approach by building links among observable item nodes themselves, bypassing concept links (see for example Dowling and Hockemeyer, Kambouri et al., Desmarais et al., 2001, 1994, 1996). They emerge from the work of Falmagne et al. (1990) and Doignon and Falmagne (1999) on the theory of knowledge spaces. Our own work on Partial Order Knowledge Structures (POKS) (Desmarais et al., 1996; Desmarais and Pu, 2005) falls under this line of research as well.

Item to item structures are good candidates for learned student models because their nodes are observable, in contrast to concept nodes. However, systems that make use of student models, like intelligent tutoring systems, deal at the concept level, not at the item level. Item to item structures thus need a mean to bridge items to concepts. There exist many means and we will return to them later in section 7. For now, let us focus on the problem of building item to item graph structures and performing an assessment at that level.

3.1. Knowledge Spaces

Item to item structures are based on a cognitive modeling theory named knowledge spaces (Doignon and Falmagne, 1999). The theory of knowledge spaces asserts that knowledge items, i.e. observable elements that define a knowledge state such as question items, are mastered in a constrained order. A knowledge state is simply a subset of items that are mastered by an individual and the knowledge space determines which

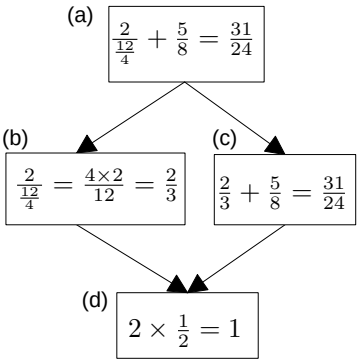


Figure 1. A simple knowledge space composed of 4 items ($\{a, b, c, d\}$) and with a partial order that constrains possible knowledge states to $\{\emptyset, \{d\}, \{b, d\}, \{c, d\}, \{b, c, d\}, \{d, b, c, a\}\}$.

other state the person can move to. Viewed differently, the knowledge space defines the structure of prerequisites among knowledge items. For example, we learn to solve figure 1's problems in an order that complies with the inverse of the arrow directions. It follows from this structure that if one masters knowledge item (c), it is likely she will also master item (d). Conversely, if she fails item (c), she will likely fail item (a). However, item (c) does not significantly inform us about item (b). This structure defines the following possible knowledge states (subsets of the set $\{a, b, c, d\}$):

$$\{\emptyset, \{d\}, \{c, d\}, \{b, d\}, \{b, c, d\}, \{a, b, c, d\}\}$$

Other knowledge states are deemed impossible (or *unlikely* in a probabilistic framework).

Formally, it can be shown that if the space of individual knowledge states is closed under *union*, then that knowledge space—the set of all possible knowledge states—can be represented by an AND/OR graph (Doignon and Falmagne, 1999). In other words, if we combine two individuals' knowledge states, then that combined knowledge state is also plausible (i.e. part of the knowledge space). However, knowledge spaces are not closed under *intersection*, meaning that if we take the common knowledge items between two individuals' knowledge states, then we can obtain an invalid knowledge state. This phenomenon occurs when a knowledge item has two alternative prerequisites. For example, one individual might learn to add two fractions by first transforming

them into a common denominator, whereas someone else might have learned to transform them into decimal form first, and then back into a rational form. If each of them ignores the other individual's method, then the intersection of their knowledge states yields a state with the mastery of the fraction addition problem with none of the other two prerequisite knowledge items being mastered.

For cases where such alternative methods of solving a problem exist, then the alternative prerequisite items will be linked to the problem with an OR relation (eg. $A \rightarrow B \vee C$), indicating that only one of the prerequisite is required (B OR C) is a required prerequisite of A. If all prerequisite were required, they would be linked with an AND relation. See Carmona et al. (2005) for an example of a BN used in modeling a structure of AND/OR relations between prerequisites in the domain of arithmetic.

For our purpose, we make the assumption/approximation that knowledge spaces are closed under *union* **and** *intersection* and ignore the possibility of representing alternate prerequisite knowledge items. We refer to this variant as *partial order knowledge structures*, or POKS. Such structures can be represented by a DAG (Directed Acyclic Graph)[1], such as the one in figure 1 because we further impose the assumption of closure under intersection (see Desmarais et al., 1996). This assumption allows a considerable reduction the space of knowledge states. It greatly simplifies the algorithms for inducing a knowledge structure from data and reduces the amount of data cases required.

It can be seen that the theory of knowledge spaces and its POKS derivative make no attempt to structure knowledge in a hierarchy of concepts or any other structure containing latent variables (often called *latent traits*). The knowledge state of an individual is solely defined in terms of observable evidence of skills such as test question items. Of course, that does not preclude the possibility to re-structure knowledge items into higher level concepts and skills. In fact, this precisely is what a teacher does for developing a quiz or an exam, for example.

3.2. Partial Order Knowledge Structures and Bayesian Networks

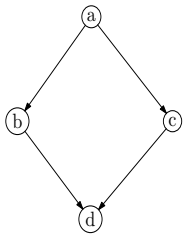
Although POKS networks like the one in figure 1 can be conveniently represented graphically by a DAG that resembles to a BN, the semantics of links is different. BN directed links are often used to represent causal relationships (although they can represent many kind of probabilistic relationship) and the structure explicitly represents conditional

[1] See Doignon and Falmagne (1999) for a formal proof and thorough analysis of the formal properties of knowledge spaces.

independence between variables. A Knowledge space directed link is similar to a logical implication relation, but it represents a prerequisite, or, to use the knowledge spaces terminology, a *surmise* relation. For example, if we have a surmise relation $A \succ B$, it implies that the mastery of B will precede the mastery of A, and thus if a student has a success for A, that student is likely to have a success for B. Moreover, its structure represents a partial ordering, namely the order in which items are likely to be learned.

That difference in the semantics of links has a number of implications. For one, the closures under *union* and *intersection* of POKS implies that, given a relation $A \rightarrow B$, the absolute frequency of people who master a knowledge item A will necessarily be smaller or equal to the frequency of B. This conclusion does not hold for the case of general Bayesian networks. For example, take the following BN that has the same topology as figure 1's (a BN taken from Neapolitan, 2004):

(a) smoking history
(b) bronchitis
(c) lung cancer
(d) fatigue



It is clear that *smoking history* (a) can be a much more frequent state than *lung cancer* (c) and *bronchitis* (b). In POKS, the frequency of (a) cannot be higher than that of (b) and (c). It is also obvious that, whereas the occurrence *lung cancer* could decrease the probability of *bronchitis* by "discounting" that later cause as a plausible explanation for *fatigue*, discounting does not play a role in the case of knowledge structures (eg. observing figure 1's (c) would not decrease the probability of (b)).

In short, many interactions found in general BN do not occur in POKS. We conjecture that this reduction in the space of possibilities that characterizes POKS, namely the closure under *union* and *intersection*, can be used to infer knowledge structures with algorithms that rely upon stronger assumptions and more constrained search spaces than for the more general BN models.

In fact, structural induction techniques tailored to the Knowledge structures and the POKS frameworks have been devised by a number of researchers. For example, Kambouri et al. (1994) introduced a semi-automated algorithm to construct knowledge structures. They developed an application that combines the use of empirical and an interactive question-answer process with domain experts to successfully

construct knowledge structures. Their approach allows the construction of knowledge structures closed under *union* only, which implies it can represent alternative prerequisites. However, the construction process requires human intervention and cannot be considered as automated learning.

In the current study, we focus on the construction of item to item structures solely from automated learning approaches. The next section describes the two approaches we studied to construct item to item structures, namely a generic BN learning and inference scheme, and a constraint-based structural learning approach to induce POKS item to item structures that relies on the local independence assumption.

4. Learning Item to Item Structures

The topology of item to item structures can be fairly intertwined and complex. Inducing such a structure is a difficult task to perform manually. It entails determining the order of mastery among knowledge items. If the set of knowledge items is large, over a few tens of items for example, our own experience is that this task can be very tedious and error prone.

Thus, finding means of learning the item to item knowledge structures from empirical data is imperative. We study two means of learning item to item structures:

- Bayesian Network structural learning;
- a POKS (Partial Order Knowledge Structure) learning algorithm.

Each approach is discussed below. Experiments to compare their respective performance for predicting item responses outcomes is reported later.

4.1. Bayesian Network Structural Learning for Item to Item Structures

In spite of the semantic differences between the links of a BN and those of an item to item structure like Figure 1's, the relations of both structures can be thought of as probabilistic implications between nodes. Both can represent evidence that influences the probabilities of neighboring nodes taking on values of true or false, in accordance to a Bayesian framework. It follows that any BN structural learning algorithm is a reasonable candidate for learning item to item structures.

We conducted a study on learning item to item BN structures with the K2 (Cooper and Herskovits, 1992) and PC algorithms (Spirtes

et al., 2000). These algorithms are regularly used in the BN structure learning literature.

K2 The general principle of the K2 algorithm is to maximize the probability of a given topology given observed data. It uses a greedy search algorithm over the space of network topologies (Cooper and Herskovits, 1992). The search is constrained by a given initial node ordering pattern to reduce the search space. For our experiments we use the topological order obtained from running the Maximum Weight Spanning Tree (MWST) algorithm by (Chow and Liu, 1968) to derive a network topology, and by extracting a topological order from this structure. François and Leray (2003) has shown that the initial DAG obtained by the MWST is an effective replacement to a random ordering. We also used Cooper and Herskovits (1992) original Bayesian metric to score the structures.

PC In contrast to searching the space of network topologies using a global Bayesian metric to score the topologies, the PC algorithm (Spirtes et al., 2000) falls into the *constraint-based structural learning* approach. It uses local conditional independence tests between a set of nodes to determine the network topology. Heuristic search consists in adding and deleting links according to the results of the independence tests and the search strategy. According to Murphy (2001), the PC algorithm is in fact a faster but otherwise equivalent version of the IC algorithm from Pearl and Verma (1991).

The results of applying these techniques over three data sets is reported in section 6.4.

4.2. POKS Structural Learning

The second approach for inducing the relations among items is based on Desmarais et al. (1996). We refer to it as the POKS induction algorithm. This approach to learning can be considered a constraint-based structural learning approach since it uses conditional independence tests to determine the structure. The POKS induction algorithm relies on a pairwise analysis of item to item relationships. The analysis attempts to identify the order in which we master knowledge items in accordance to the theory of knowledge spaces (Doignon and Falmagne, 1999) but under the stronger assumption that the skill acquisition order can be modeled by a directed acyclic graph, or DAG.

The tests to establish a relation $A \rightarrow B$ consists in three conditions for which a statistical test is applied:

$$P(B|A) \geq p_c \tag{1}$$

$$P(\overline{A}|\overline{B}) \geq p_c \tag{2}$$

$$P(B|A) \neq P(B) \tag{3}$$

Conditions (1) and (2) respectively correspond to the ability to predict that A is true given that B is observed true (*mastered*), and the ability that B is false (*non-mastered*) given that A is false. The third condition verifies that the conditional probabilities are different from the non conditional probabilities (i.e. there is an interaction between the probability distributions of A and B). The first two conditions are verified by a Binomial test with parameters:

p_c the minimal conditional probability of inequalities (1) and (2),

α_i the alpha error tolerance level.

The conditional independence test is verified by the Fisher exact test and the χ^2 test can also be used. See Desmarais et al. (1996) or Desmarais and Pu (2005) for further details about the parameters.

For this study, p_c is set at 0.5. Condition (3) is the independence test verified through a χ^2 statistic with an alpha error $\alpha_i < 0.2$. The high values of alpha errors maximize the number of relations we retain.

5. Inferences

Once we obtain an item to item structure, an assessment of the probability of success over all items can be computed from partial evidence. In other words, we wish to evaluate the validity of the two frameworks over their item outcome predictive ability. We do not attempt to assess the validity of the actual item to item structures themselves because we have no mean to determine their respective true structure. In fact, that issue belongs to the field of cognitive science and was already thoroughly investigated by Doignon and Falmagne (see Doignon and Falmagne, 1999) and a number of other researchers. Our interest lies in the predictive power of the models which is measured by their ability to perform accurate assessment.

5.1. Inference in BN

For the BN structure, there exist a number of standard and well documented algorithms (see, for eg., Neapolitan, 2004). We use the junction-tree algorithm (Jensen, 1996) which preforms an exact computation of posterior probabilities within a tree whose vertices's and derived from a triangulated graph, which is itself derived from the DAG in the BN.

5.2. Inference in POKS

For the POKS framework, computation of the nodes' probabilities are essentially based on standard Bayesian posteriors under the local independence assumption. We use a slightly different version than the one proposed in Desmarais and Pu (2005) and it is described below.

Given a relation $E \rightarrow H$, where E stands for an evidence node (parent) and H stands for a hypothesis node (child), the posterior probability of H is computed by using the odds likelihood version of Bayes' Theorem:

$$O(H \,|\, E) \;=\; O(H)\, \frac{P(E \,|\, H)}{P(E \,|\, \overline{H})} \tag{4}$$

$$O(H \,|\, \overline{E}) \;=\; O(H)\, \frac{P(\overline{E} \,|\, H)}{P(\overline{E} \,|\, \overline{H})} \tag{5}$$

where $O(H)$ is the prior odds ratio and $O(H|E)$ represents the odds of H given evidence of E, and assumes the usual odds definition $O(H|E) = \frac{P(H|E)}{1-P(H|E)}$.

In order to make inference from combined evidence sources, the knowledge structure inference process makes the local independence assumption. Given that assumption, the computation of a joint probability of evidence nodes, $E_1, E_2, \ldots, E_i$, and the hypothesis node, H, is a straightforward product of likelihoods. For example, assuming that we have n number of relations of the form $E_i \rightarrow H$, then it follows from this assumption that:

$$P(E_1, \ldots, E_n \,|\, H) = \prod_i^n P(E_i \,|\, H) \tag{6}$$

From equation (6), it follows that the probability update of H given $E_1, ..., E_n$ can be written in following posterior odds form:

$$O(H \,|\, E_1, E_2, \ldots, E_n) = O(H) \prod_i^n \frac{P(E_i \,|\, H)}{P(E_i \,|\, \overline{H})} \tag{7}$$

In case the evidence is negative for observation i, then the ratio $\frac{P(\overline{E_i} \,|\, H)}{P(\overline{E_i} \,|\, \overline{H})}$ is used.

The local independence is a strong assumption that is a characteristic of the naive Bayes framework. It greatly simplifies the amount of data required to calibrate conditional probabilities. Although this assumption is very likely violated to a certain degree in many cases,

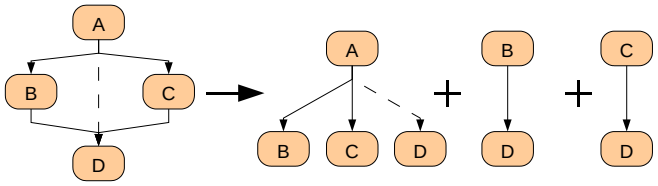


Figure 2. Correspondence of a DAG to a set of single layer networks used for forward inferences. Backward inferences follow arrows in the opposite direction.

it was shown to be robust in many situations (see Domingos and Pazzani, Rish, Friedman et al., 1997, 2001, 1997).

5.2.1. *From DAG to Single Layered Networks*
In the current study, we do not use transitive/recursive propagation to perform inference based on partial evidence as was done for previous studies with POKS (Desmarais and Pu, 2005; Desmarais et al., 1996). For example, if we have $A \rightarrow B$ and $B \rightarrow C$, no probability update is performed over C upon the observation of A, unless a link $A \rightarrow C$ is explicitly derived from the data. However, we rely on the fact that if we have strong surmise relations $A \rightarrow B \rightarrow C$, then we would also expect to find $A \rightarrow C$. This follows from the fact that in a knowledge space closed under union and intersections, surmise relations are equivalent to logical implications.

This principle is illustrated in figure 2. A simple POKS topology can be transformed into three single layered networks. The dotted line in the partial order would normally be derived from data if the network contains strong surmise relations.

The departure from the original POKS framework (Desmarais et al., 1996) makes the model simpler. It avoids the definition of a scheme to propagate partial evidence: propagating evidence from A to C in a structure like $A \rightarrow B \rightarrow C$, for example. Given that we expect partial evidence inferences to result in direct, transitive relations, the results are expected to be very similar. This was confirmed in our own experimental results that show that the performance is very close between the two alternatives illustrated by figure 2.

5.3. Symmetric Relations

If the sample data conformed perfectly with the assumption of closure under *union* and *intersection* and there were no knowledge items that are strictly equivalent, the result of the POKS network induction algorithm would yield a DAG. If two knowledge items are actually

equivalent in terms of prerequisites and difficulty, then the network would contain symmetric relations between equivalent knowledge items. Nodes linked with symmetric relations could then be collapsed into a single node, transforming the graph into a DAG and considering the two merged items as two equivalent instances of a single knowledge item.

However, the reality is not so simple. In practice, many nodes have symmetric relation derived according to the induction algorithm of section 4.2, but the items do differ in terms of prerequisites and difficulty level. The more tolerant the induction process is (i.e. the lower the values of p_c and the higher α_i is in inequalities (1), (2), and (3)), the more symmetric relations with non equivalent nodes will occur. This will be reflected by symmetric relations with different values of $O(H|E)$ (see section 5.2). For example, consider a symmetric relation, $A \leftrightarrow B$. If both items are equivalent, then $O(A|B) \approx O(B|A)$ and both odds will have very high values. But if they are not equivalent, a symmetric relation can still occur while $O(A|B) \not\approx O(B|A)$.

The consequence of having symmetric relations is that the actual structure derived by the POKS induction algorithm is not a strict DAG. In the current experiment, cycles introduced by symmetric relations have no impact on the inference algorithm because we do not propagate over partial evidence (transitive relations). However, when using an algorithm that recursively propagates evidence through transitively connected nodes, such as in Desmarais and Pu (2005) and Desmarais et al. (1996), care must be taken over these cycles. A simple and standard rule is to stop propagation when a node has already been updated with a given evidence source.

Note that cycles such as $A \rightarrow B$, $B \rightarrow C$ and $C \rightarrow A$ could not occur other than by following a symmetric relation. See Appendix A for a formal proof.

6. Predictive Comparison of the BN and POKS Structural Learning Approaches

The BN and POKS structural learning approaches of item to item structures are compared over their ability to predict item response outcome. We use data from real tests to conduct simulations and measure the performance of each approach for predicting the outcome over the full set of item answers from a subset of observed answers. This validation technique is identical to the ones used by Vomlel (2004) and Desmarais and Pu (2005).

6.1. Simulation Methodology

The experiment consists in simulating the question answering process with the real subjects. An item is chosen and the outcome of the answer, success or failure, is fed to the inference algorithm. An updated probability of success is computed given this new evidence. All items for which the probability is above 0.5 are considered mastered and all others are considered non-mastered. We then compare the results with the real answers to obtain a measure of how accurate the predictions are. The process is repeated, starting from 0 item administered, until all items are "observed". Observed items are bound to their true value, such that after all items are administered, the score always converges to 1.

The simulations replicate a context of computer adaptive testing (CAT) where the system chooses the question items in order to optimize skills assessment. This context is typical of study guide applications, where a quiz is administered prior to providing pedagogical assistance (Falmagne et al., 2006; Dösinger, 2002). However, the choice of question may not entirely be driven by the need to optimize skills assessment, but also by an adaptive pedagogical strategy such as in Heller et al. (2006), for example.

For this experiment, the choice of the question to ask is determined by an entropy reduction optimization algorithm. The same algorithm is used for both the BN and POKS frameworks (for details on this algorithm, see Vomlel, 2004, Desmarais and Pu, 2005). Essentially, the choice of the next question to administer corresponds to the one that reduces the entropy of a set of network nodes. The algorithm will choose the item that is expected to reduce entropy the most. Items with very high or low probability of success are generally excluded because their expected entropy reduction value will be low.

6.2. Data Sets

The data sets are taken from three tests administered to human subjects :

1. **Arithmetic test.** Vomlel (2004) gathered data from 149 pupils who completed a 20 question items test of basic fraction arithmetic. This data has the advantage of also containing independent concept assessment which we will return to when assessing the approaches' ability to predict concepts.

2. **UNIX shell command test.** The second data set is composed of 47 test results over a 33 question items test on knowledge of

Table I. Data sets

Data set	nb. items	nb. cases			Average success rate
		Training	Test	Total	
Arithmetic	20	100	49	149	61%
Unix	33	30	17	47	53%
French	30	30	12	42	50%

different Unix shell commands. The questions range from simple and essential commands (eg. *cd*, *ls*), to more complex data processing utilities (eg. *awk*, *sed*) and system administration tools (eg. *ps*, *chgrp*).

3. **French language test.** The third data set consists of a standard test of French language administered by the Government of Canada. The test is actually composed of 160 items but, because the BN software for our experimentation cannot handle that many items, we randomly selected 30 of these items. The selection is in fact a stratified sampling to ensure that items of all difficulty levels are chosen. The number of test data cases is 42.

For each data set, a portion of the data is used for training and the remaining ones for testing. Table I provides the size of the training and testing sets along with the average success rate of each test.

For each corpus data set, six training and test sets were randomly sampled and a simulation run is performed with each of these six sample pairs. All performance reports represent the average over all six sample runs.

6.3. Simulation Parameters

The BN and the POKS structural learning approaches require that a number of parameters be set.

The K2 and PC algorithms are tested for the BN structure learning. The BN parameters for both algorithms were initialized with Dirichlet uniform priors, which correspond to Beta priors in the case of binomial variables.

The PC algorithm must be given a value for the interaction test significance level. We used a value of 0.2, which is the same as for POKS α_i parameter in inequalities (1) and (2) (see below).

As mentioned, the K2 algorithm is fed with a node ordering obtained from running the Maximum Weight Spanning Tree (MWST) algorithm

by (Chow and Liu, 1968) and by extracting a topological order from this structure. François and Leray (2003) has shown that the initial DAG obtained by the MWST is an effective replacement to a random ordering. We also used Cooper and Herskovits (1992) original Bayesian metric to score the structures.

We used Ken Murphy's BNT Matlab package for learning the BN structures of all the experiments conducted (`http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html`). Note that it was not possible to test the PC algorithm for the Unix and French tests because of resource limitations for Matlab©.

The POKS structural learning algorithm also involves parameters that need to be set: the minimal conditional probability level, p_c, and the alpha error tolerance level, α_i (see section 4.2). The two parameters' values are:

$$p_c = 0.5 \text{ and } \alpha_i = 0.2$$

These values were also used in Desmarais et al. (1996) and they are generally appropriate when the number of nodes is below 50.

6.4. Learned Structures

Over all six randomly sampled sets, the POKS structural learning algorithm created structures that, for the arithmetic data set, contains between 181 and 218 relations, of which 117 to 126 are respectively symmetric, for an average between 9.1 to 10.9 links per node. For the Unix data set, the number of relations varies between 582 and 691, with the number of symmetric relations that varies between 348 and 297. The average relations per node varies between 17.6 to 20.9. The structure of the Unix data set is thus much more populated with an average link per node about twice that of the arithmetic test. These structures are too dense to be shown graphically here.

For the BN structural learning results, figure 3 displays the first two structures learned with the K2 algorithm. It can be seen that the topology differs significantly between the two networks shown in this figure. In general, about only half of the relations are common between BN from two samples. However, and as mentioned, we do not focus on the actual topologies in this study but solely on the ability of the induced structures to perform accurate inferences.

6.5. Computational resources

Processing time for learning differs substantially between the different structural learning algorithms and for performing probability updates

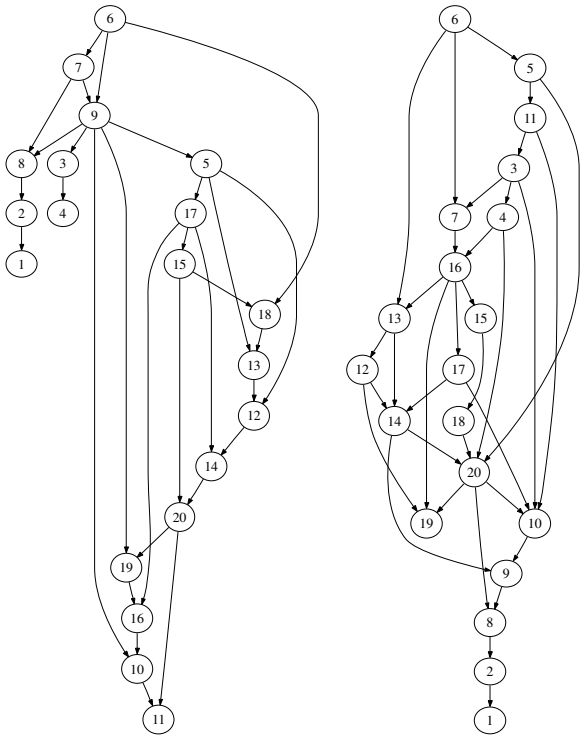


Figure 3. Two examples of BN structures learned with the K2 algorithm.

over the three data sets. Table II reports the different times for computing structure learning and for running a single subject simulation. A single simulation run consists in $q \times q/2$ probability updates, where q is the number of items for the data set.

For structure learning, the PC algorithm is the slowest with a time of about three hundred times greater than the K2 algorithm. The K2 algorithm is about 10 times slower than POKS for all data sets.

For probability updates, POKS is also the fastest by a factor of around 400 to 1000 times compared to the K2 algorithm. PC is slower than K2 by a factor of around 6.

The simplicity of the POKS algorithm is thus reflected for both structure learning and probability updates. The difference can reach three orders of magnitude for probability updates which can have significant practical implications in an operational context.

6.6. Prediction Score

The performance measure for item prediction corresponds to the number of correctly classified item mastery. If the probability of an item is

Table II. Time computations (in seconds)

	PC	K2	POKS
Arithmetic:			
single subject simulation	275	43	0.10
structure learning	1080	3.8	0.39
Unix:			
single subject simulation	na	2.7	0.12
structure learning	na	13.2	1.08
French language:			
single subject simulation	na	58	0.14
structure learning	na	9.9	0.81

above 0.5, then it is considered *mastered*. If the student's real answer matches the estimate, then we consider that the prediction is accurate. It is inaccurate otherwise. Once an item is observed, then it is by definition accurately assessed and thus the performance after all items are observed converges to 1.

The performance is averaged over all test subjects and all six random samples. The formula for computing the accuracy of the prediction after each observed item is:

$$\text{Accuracy} = \frac{\sum_k^n \frac{\sum_j^m \frac{\sum_i^r M_{ijk}}{r}}{m}}{n} = \frac{\sum_k^n \sum_j^m \sum_i^r M_{ijk}}{rmn}$$

where r is the total number items in the test, m is the number of test subject cases (17 for the Unix test and 49 for the arithmetic test), and n is the number of random sample runs of the simulation (6). M_{ijk} represents the item outcome prediction to item i by subject k for the simulation run j. It is 1 if the prediction is correct or if it is an observed item, and 0 otherwise.

6.7. Results

Figure 4 reports the simulations results. Each curve represents the accuracy score along with a 90% confidence interval computed over the six simulation runs. It shows that, for all three data sets, the POKS algorithm yields more accurate predictions of item outcome than the BN algorithms. Although the difference is only a few percentage points,

it is relatively substantial. For example, after 10 items, the difference between the BN and POKS for the Unix data set is about 92.5% compared to 89.4% for the K2 algorithm. Although this represents a 3% difference in absolute values, it should be regarded relative to error reduction. In terms of the remaining error, it represents a 30% relative reduction such that the system would reduce the number of wrong decisions from 10 in 100, to 7. Viewed from a different perspective, it means that the accuracy reached by POKS after 10 of the 33 item Unix test is only reached after about 14 items for the BN K2 algorithm. In a context where, for example, we need strong confidence that a specific item is mastered and avoid making wrong decisions from an incorrectly assessed item, the difference in reliability can result in substantially fewer items administered.[2]

Looking at the confidence intervals, we note that they are around twice as large for the K2 algorithm than for POKS. This can also have a practical impact when determining the certainty of a decision and whether we need more evidence or not.

We also note that the PC algorithm performs better than the K2 algorithm in the Arithmetic test, partly due to more accurate priors, although the difference quickly vanishes after 2 items observed and becomes is insignificant.

Table III reports the statistical significance for the different comparisons between the algorithms and the data sets. A paired Student-t test is performed after each observed item on the x-axis. Each data point in the test conditions represents the accuracy averaged over subjects for a given simulation run. There are 6 runs, such that the number of degrees of freedom for the Student-t test is $6-1=5$. All comparisons with the POKS algorithms show a significant advantage at the $p < 0.05$ level (or lower), except around some of the extremes. The difference between the K2 and the PC algorithms are not significant except at the beginning where the priors favor the PC algorithm.

6.8. Discussion

The better performance of the POKS approach over a BN approach may appear surprising, since both schemes rely on the Bayesian framework and the POKS approach makes stronger assumptions than the BN approach. However, this is not an exception. POKS can be considered as part of the naive Bayes family of models, because it makes the

[2] Refer to the section on accuracy and reliability in the introduction where we emphasize to the need for high accuracy when the cost of taking a wrong decision is high. Models that can yield reliable diagnostics are thus very important in such contexts (see also Horvitz, 1999).

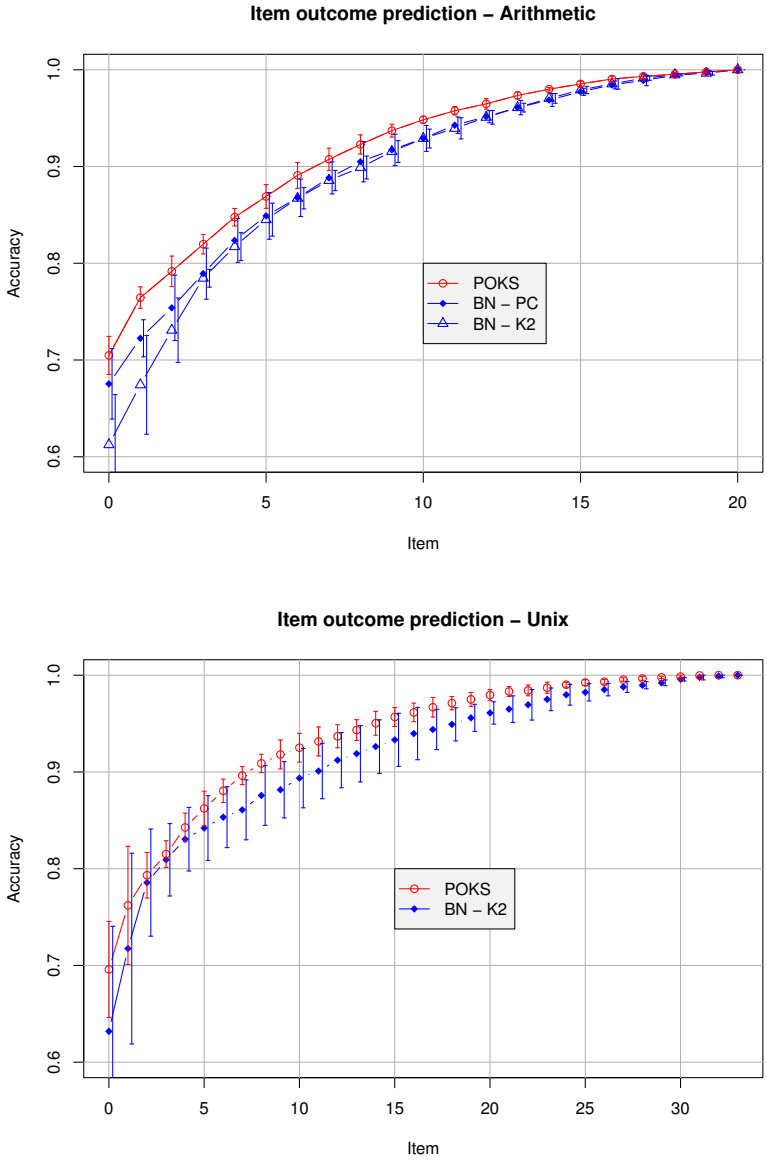


Figure 4. Item prediction performance. The graphs report the accuracy of item outcome predictions for two data sets. A third one is reported in the next figure (Figure 5). Each line represents an average over 6 simulation sample runs and, for each data point, a 90% confidence interval over subjects is shown. The simulation runs consist of 49 test cases for the Arithmetic data set, 18 for Unix and 12 for the French data set shown in figure 5.

Table III. Subject paired Student-t tests for all data sets (N=6).

Items	Arithmetic			UNIX	French
	POKS-K2	POKS-PC	PC-K2	POKS-K2	POKS-K2
0	-	**	**	-	*
1	***	**	*	-	*
2	*	**	-	-	*
3	*	***	-	-	*
4	**	**	-	-	**
5	*	*	-	-	**
6	**	*	-	-	**
7	**	**	-	**	**
8	**	***	-	**	**
9	**	**	-	**	**
10	**	***	-	**	**
11	**	***	-	**	**
12	**	***	-	*	**
13	***	***	-	*	**
14	**	**	-	**	**
15	**	*	-	*	**
16	*	-	-	*	**
17	-	-	-	*	**
18	-	-	-	**	**
19	-	-	-	**	**
20				**	*
21				**	**
22				*	**
23				*	**
24				*	-
25				*	-
26				*	-
27				*	-
28				**	-
29				**	-
30				*	
31				-	
32				-	

*** $p < 0.001$
** $p < 0.01$
* $p < 0.05$
- $p > 0.05$.

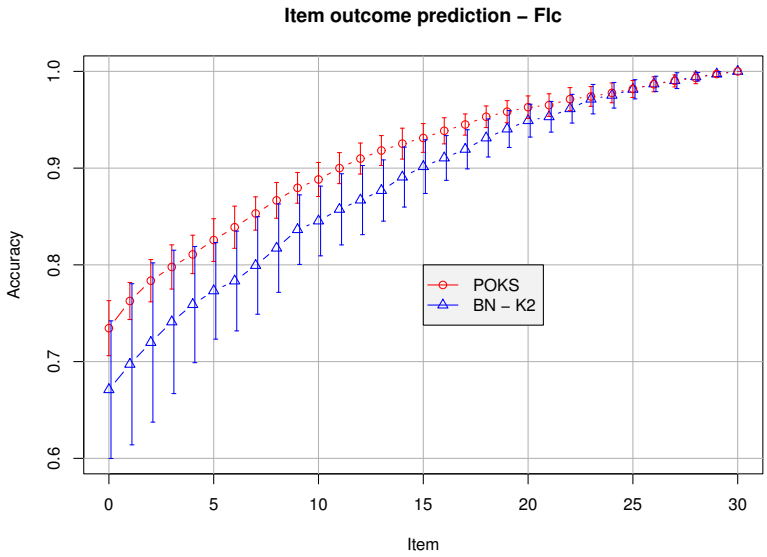


Figure 5. Item prediction performance for the French data set. See Figure 4's caption for details.

independence assumption among *features* (the observed nodes in our context). In spite of this simplifying assumption, naive Bayes models were shown very effective in a large array of contexts (Domingos and Pazzani, 1997; Rish, 2001). The context of POKS may also be the case. We return to this issue in the overall discussion.

7. Combining Concept Nodes and Item to Item Structures

We argued earlier that systems that can make use of student models, be it adaptive hypertext, intelligent tutoring system, or study guides, need to work at the level of concepts, not at the item level. We now turn to the issue of using item to item structures for assessing concept mastery.

7.1. Rule Space, Weighted Means

Simple techniques such as Tatsuoka's Rule Space or a weighted means are valid alternatives to assess concepts.

Tatsuoka (1983) introduced the concepts of Rule space and Q-matrix. Each "rule" (concept or skill) that is considered a required condition for the success of each test item is indicated in a matrix

of rules by items. A probabilistic version of this framework actually corresponds to the approach of VanLehn et al. (1998) mentioned above.

An alternative to the Q-matrix is to decompose the mastery of a given concept as a weighted mean of items, much in the same manner as every teacher does when points are allotted to different test items in a exam. That approach has the advantage of being readily understood by teachers who frequently go through this process of determining which test items assess which concepts or topics.

These two approaches are the most simple and likely means of using an item to item assessment. It cold be used to *augment* the initial set of observations by inferring likely mastered and non-mastered items. To the extent that the inferences are accurate, it would necessarily improve the concept assessment as well. We return to this mechanism below.

7.2. Previous Work with BN

Whereas the above methods are very simple and do not really make use of relationships between concepts to make inferences, the BN approach does and this constitutes an attractive advantage over the simpler approaches.

VanLehn et al. (1998) investigated some variations on Pearl's noisy-And model (Pearl, 1988) to link observable items to concepts, and found them effective with simulated student test data. Whereas this technique represents a means to introduce evidence from items into a BN in the absence of the required conditional probability tables, Millán et al. (2000) introduces a means to fill such tables in the absence of sufficient data. They used a combination of expert judgments and IRT's logistic function to parametrize the conditional probabilities between test items and concepts. All these efforts are means to link observed items to a BN structure of concepts. We now turn to an architecture that augments observed items with an item to item approach that can either replace or complement these techniques.

7.3. Augmenting the Observed Evidence Set

Assuming a link from observable evidence to concepts, we can use the item to item model to augment the initial set of observed evidence and feed this augmented evidence set to the concept level model. We already hinted on such technique above with Q-matrices and weighted sums, where the gain from the augmented inferences is obvious, but it is also feasible with more sophisticated approaches. For example, an item to item model could feed a BN with an augmented response set that complements the information used by the BN at the concept level. To the extent that the item to item model provides an accurate assessment,

we would expect that the assessment at the concept level would also be improved. This approach is investigated in the next section.

8. Experimental Assessment of a Hierarchical BN Combined with an Item to Item Structure

The previous section describes a few potential techniques to combine a BN with an item to item structure. We now experiment with one such technique, the augmented set of evidence technique outlined in section 7.3. We specifically wish to verify if an item to item structure like POKS can yield information that could improve the accuracy of concept and item assessment over a BN defined without item to item links.

In order to investigate this question, we use the data from Vomlel (2004) which not only contains the 20 test items that served for the experiment reported in section 6, but it also includes 20 concept nodes that were independently assessed by human experts. This independent assessment allows cross-validation measures. Figure 6 illustrates the BN found in Vomlel (2004).

This data allows a comparison of predictions at two levels:

1. *item predictive accuracy*: evaluate the performance for predicting actual responses to individual items;
2. *concept predictive accuracy*: evaluate the performance for assessing concept mastery based on an independent source;

8.1. The BN Model

Figure 6's BN model of this experiment is a quasi-hierarchical decomposition of basic concepts in fraction arithmetic. Item nodes are represented by leaves of this structure.

Vomlel experimented with a number of BN models to determine each model's ability to predict the actual question item success (item predictive accuracy) and concept mastery (concept predictive accuracy). A total of 9 BN models were tested, from a naive Bayes model to structures with different number of hidden nodes and structures that were partly constructed by hand. We report the results of the best performing one, which is the one reproduced in figure 6. This structure was in part constructed by an expert and by structural learning with the HUGIN software.

For assessing the mastery of concepts, Vomlel used an independent source: expert judgment on the mastery of each concept based on the

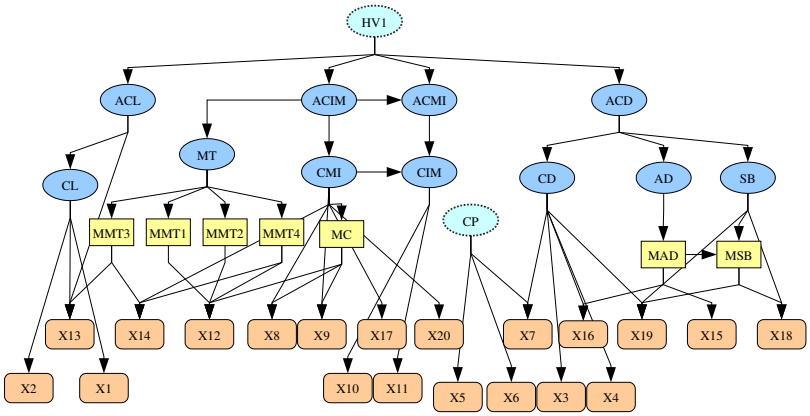


Figure 6. Bayesian network from Vomlel (2004). The model contains 20 question item nodes represented by leaf nodes. Other nodes represent concepts (oval shape) or misconceptions (rectangle shape). There are two hidden nodes, HV1 and CP.

specific answer pattern to each 20 question items. That data allows the training of the BN as if concepts, that are usually hidden nodes in a BN, could actually be observed. This situation is atypical, since we generally do not have the luxury of "observing" concept mastery and of training a model with such data, but it conveniently allows us to do an experimental comparison of the performance of different models to predict concept mastery.

Vomlel (2004) used the independent concept mastery assessment data to calibrate the conditional probabilities between items and concepts. The concepts themselves become observable nodes for the training phase. Training is performed over all subjects except one: the subject used in the simulation. This simulation method allows the use of $N-1$ data cases, while avoiding the bias in using the same data for training and validation.

8.2. Concept and Item Predictive Accuracy

We use figure 6's BN to make predictions at the item and concept level by replicating the experiment from Vomlel (2004). The item selection choice is based on the entropy reduction algorithm, akin to the algorithm used for the item to item experiment of section 6, except that entropy reduction is geared to reducing concepts and item entropy

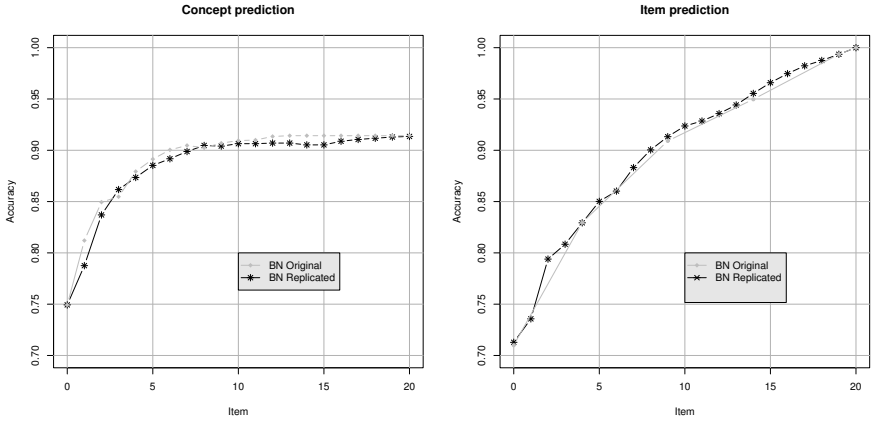


Figure 7. Comparison of Vomlel's (2004) original results with our replicated results.

simultaneously. The results of our replication of Vomlel's experiment is reported in figure 7.

Both the original and replicated results are relatively similar, as expected since we use the same generic algorithms. Implementation details and the stochastic processes involve can explain the small difference. Figure 6's BN is specified and the conditional probabilities are determined through the EM algorithm, akin to Vomlel (2004). The EM algorithm is used because the concept assessment data contains approximately 9% of missing values. The junction-tree inference algorithm is used in both our experiment and Vomlel's.

The performance measure is the same as the one explained in section 6.6.

8.3. Combination Algorithm

We assess the effectiveness of the evidence augmentation scheme outlined in section 7.3 for combining the POKS item to item structure with figure 6's BN. More specifically, we use POKS as a filter to augment the actual number of observations fed to the BN. This process is illustrated in figure 8. Assuming a set of observed responses S, POKS infers a set of additional responses, S'. The original set, S, is thereby augmented by the inferences from POKS, S' and the set of evidence fed to the BN represents the union of S and S'. This process is repeated for every new observation, from 0 to all 20 items.

In order to determine that an item is considered inferred by POKS, a threshold is used, δ. Every item for which the probability of mastery of POKS is greater than $1 - \delta$ is considered mastered, whereas items with a probability smaller than δ are considered non-mastered.

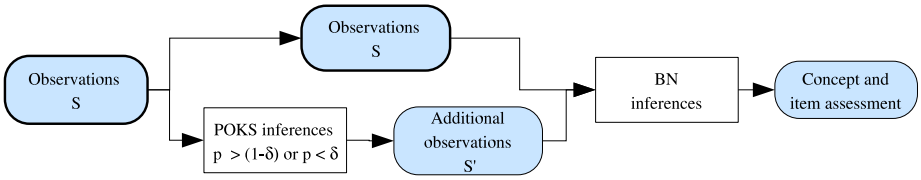


Figure 8. Combination algorithm of POKS with BN.

8.4. Results

The simulation results of the combination algorithm of POKS and the BN are reported in figure 9. A threshold $\delta = 0.1$ is used for this experiment. This value provided the best results, although there were only small differences between about 0.3 and 0.95. We did not explore asymmetric thresholds for success and failures.

The four data lines reported represent three variations over the item selection strategy. The graph also includes the original simulation result without any augmentation:

POKS: original simulation results of the POKS item prediction performance.

BN+POKS item entropy: augmented inferences using the item selection algorithm based on POKS item nodes entropy.

BN item entropy: BN only inferences using the item selection based solely on the BN's item, excluding concept nodes. The formula for computing entropy thereby only include item nodes.

BN global entropy: BN only inferences using the original the item selection algorithm based on global entropy (items and concepts).

Table IV and V respectively report the standard deviations and the statistical significance of the differences observed in Figure 9. The standard deviation is computed over subjects, not over simulation runs as it was for the item outcome prediction (figures 4 and 5), since the methodology here replicates that of Vomlel (2004) and consists in a single simulation where a new structure is learned for every subject by first removing her data case from the training data. The standard deviations are thus much greater as they represent the dispersion of subject scores. However, because the number of subjects is 149, the degrees of freedom of the Student-t test is 148 and, for similar differences, it results in greater statistical significance than for the 5 degrees of freedom we had in table III.

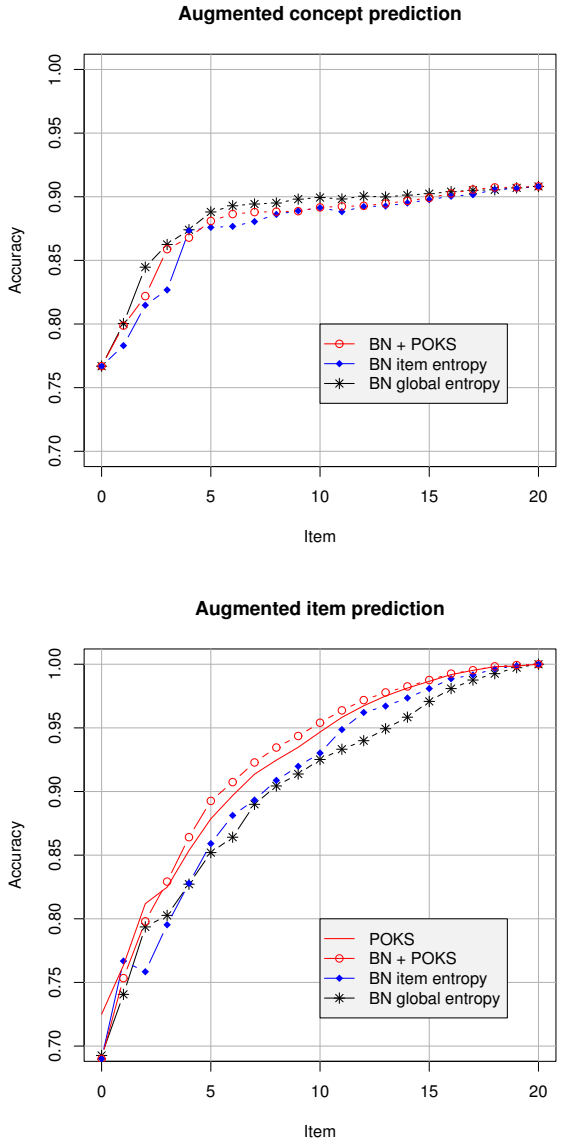


Figure 9. Results of the simulation where the POKS inferences are used to augment the observed set of items according to figure 8's diagram. A threshold value of $\delta = 0.1$ is used. Refer to the text for a description of each curve.

Table IV. Concept standard deviations and Student-t test results (N=149).

Item	Standard deviations			Student-t tests		
	(1)	(2)	(3)	(2)-(3)	(1)-(3)	(1)-(2)
0	0.177	0.177	0.177	-	-	-
1	0.170	0.107	0.130	-	-	-
2	0.109	0.093	0.122	***	**	-
3	0.117	0.085	0.088	***	-	***
4	0.106	0.102	0.093	-	-	-
5	0.087	0.093	0.096	*	-	-
6	0.088	0.095	0.097	***	-	-
7	0.088	0.094	0.095	***	-	-
8	0.087	0.089	0.095	*	-	-
9	0.087	0.087	0.097	*	*	-
10	0.085	0.086	0.094	-	-	-
11	0.083	0.091	0.094	*	-	-
12	0.079	0.091	0.095	-	-	-
13	0.080	0.091	0.094	-	-	-
14	0.083	0.093	0.090	-	-	-
15	0.085	0.092	0.088	-	-	-
16	0.085	0.090	0.086	-	-	-
17	0.085	0.089	0.084	-	-	-
18	0.083	0.083	0.082	-	-	-
19	0.080	0.083	0.082	-	-	-

(1) BN+POKS,
(2) BN item entropy,
(3) BN global entropy

*** $p < 0.001$
** $p < 0.01$
* $p < 0.05$
- $p > 0.05$

The item prediction results reveal that the highest performance is achieved when the BN inferences are augmented by the observations from the POKS item selection strategy. For the purpose of comparison with the original POKS performance, we reproduce the corresponding curve of Figure 4 and we see that the original performance was slightly lower. These results suggest that item to item structures can provide additional, complementary inference when it comes to predicting item outcome. However, the improvement is only marginally better than the POKS alone condition, suggesting the BN did not add much information, if any, to the item assessment coming out of POKS.

Table V. Item standard deviations and Student-t test results. See table III for statistical significance symbols.

	Standard deviations			Student-t tests		
Item	(1)	(2)	(3)	(2)-(3)	(1)-(3)	(1)-(2)
0	0.152	0.155	0.155	-	-	-
1	0.211	0.137	0.150	**	-	***
2	0.144	0.129	0.120	***	-	***
3	0.132	0.113	0.106	-	***	***
4	0.107	0.107	0.109	-	***	***
5	0.091	0.096	0.102	-	***	***
6	0.088	0.099	0.093	***	***	***
7	0.093	0.097	0.089	-	***	***
8	0.090	0.091	0.081	-	***	***
9	0.082	0.088	0.074	-	***	***
10	0.079	0.078	0.065	-	***	***
11	0.074	0.069	0.060	***	***	**
12	0.071	0.062	0.052	***	***	*
13	0.066	0.058	0.046	***	***	*
14	0.061	0.048	0.041	***	***	*
15	0.053	0.042	0.035	**	***	*
16	0.043	0.033	0.026	**	***	-
17	0.036	0.028	0.019	-	**	*
18	0.028	0.017	0.012	-	**	-
19	0.014	0.010	0.007	-	-	-

(1) BN+POKS,
(2) BN item entropy,
(3) BN global entropy

*** $p < 0.001$
** $p < 0.01$
* $p < 0.05$
- $p > 0.05$

The results also show an improvement for the item selection strategy based on BN item entropy, especially after 10 items where it becomes very close to the POKS+BN condition. It reveals the importance that the item strategy can have in the item outcome prediction accuracy.

Contrary to item prediction, the concept prediction accuracy reveals that all conditions are relatively similar. Surprisingly, the improvement seen for the item outcome prediction does not transfer to the concept prediction as expected.

An explanation for this result is that the BN already captured the information inferred by POKS at the item level. In other words, the item to item structure contains redundant information to the well struc-

tured BN. The BN gets no further gain by augmenting the set of initial observations with item to item structural inferences. This would also explain the marginal gain for item assessment compared to the initial POKS assessment.

However, alternative explanations are also plausible. For one, optimizing item selection at the item inference level for the POKS+BN condition can offset the assessment gain at the concept level. This might explain why BN+POKS is sometimes slightly below the BN global entropy. But that explanation may not be the one that weights the most in the results as we see below.

The other possibility is that there is no room for improvement at the concept level. The independent concept assessment may not be reliable enough to allow any approach to reach further than the 90% level, the level reached after all observations are obtained. Given that the span of accuracy for concepts only ranges between 75% and 90%, and that maximum is almost reached after only 5 question items, that leaves little room to show improvements. We would probably have to conduct the simulation with a larger set of items in order to more reliably assess concept mastery. Currently, many concepts have only 2 or 3 items from which to validate their mastery (see Figure 6) which can lead to errors due to noise in the data as well as during the independent assessment made by the experts.

We conclude that it is plausible that a BN can capture all the information found in item to item structures. However, the experiment is not conclusive and only points to further investigations. Nevertheless, we stress that for approaches such as a weighted sum over item outcome results, or a Q-matrix, it is obvious that we would expect to see a gain at the concept level assessment since these approaches do not perform any inferences at the concept level like the BN does.

9. Discussion

Learned item to item student models have the potential to provide accurate, fine-grained skill assessment without the drawbacks of requiring significant human effort and expertise. This effort could be limited to the familiar task that every teacher goes through during the elaboration of an exam: linking and weighting items with respect to a list or a hierarchy of concepts.

This study shows that item to item structures can be constructed from data and yield effective predictive item outcome models. Two approaches were investigated, namely the standard BN framework and

the POKS framework, the later of which stands closer to the naive Bayes family of models.

Simulations over three data sets show that the POKS framework generally yields better predictive results for item outcome prediction than does the general BN framework. Given the greater simplicity of POKS over a BN framework and the considerably faster algorithms for learning and inference, this can have important practical implications.

The overall prediction accuracy gain of POKS over a BN can be explained by the conclusions of Domingos and Pazzani (1997): the objective functions of BN construction and parametrization aim to optimize the likelihood of the entire data, rather than the likelihood of the class (the *item* we wish to determine *mastered* or *non-mastered* in our case) given the attributes (the *items* observed so far). That difference in aims explains why the naive Bayes approach outperforms a BN for a classification task. Given that the process of knowledge assessment presented in our study starts by predicting whether each item is mastered or non-mastered, it can be considered as a classification task.

Other explanations can be traced to the nature of the data (for eg., see Rish, 2001). We outlined the semantic differences between POKS and those of a BN, and some characteristics that knowledge structures display and that would not occur in more general conditional dependencies. We have no evidence that these characteristics are actually linked to the performance of the naive Bayes framework, but it remains an avenue to explore.

We propose some means to exploit item to item structures with simple schemes such as a Q-matrix (Tatsuoka, 1983) and weighted sums of items, where the advantage of augmenting the initial set of observed mastered and non-mastered items with inferences at the item structure level is straightforward.

However, we further investigated how to use augmented observations from an item to item structure with an existing BN model, which offer high modeling flexibility at the concept level and enjoy great recognition in the student modeling community.

The results show that, although we can improve item outcome prediction with the augmented inference scheme, the experiment showed no improvements at the concept assessment level. One explanation is that the information contained at the concept level is redundant with the augmented set of evidence provided at the item level. In other words, all of the *augmented* (inferred) item observations were actually already derived by the concept relations in the BN and no further information was provided from the item to item structure inferences.

Although this explanation is quite plausible, the results can also be explained by limitation of the data set used. We note that concepts were independently assessed by experts and may be too noisy to reflect possible improvements from the item to item level inferences. Moreover, the concept assessment quickly converges to its 90% maximum level after 5 items and leave little room for improvement. Further studies will need to be conducted in order to confirm if the structural information at the item to item level is redundant with that found in a BN and under which circumstances.

Nevertheless, the information obtained at the item outcome level would certainly lead to better assessments in frameworks such as Q-matrices and simple weighted sums of item outcome to assess concepts, because they do not perform inferences at the concept level, as a BN would. And given that we rarely have the luxury of learning a BN from independently assessed concepts in a practical setting, the context of using Q-matrices and weighted sums for concept assessment appears much more likely to occur and make the item to item augmented observations approach useful.

A number of issues remain open over the current study, one of which is how general are the findings. We already see different patterns of results between the simulations over the three data sets. It is quite plausible that some domain of knowledge, or some types of tests, may not conform to the underlying assumptions of POKS and knowledge spaces and therefore the framework would not perform as well. Similarly, the BN structural learning algorithms can display wide differences depending on the nature of the data set and the sample size (see, for eg., François and Leray, 2003). As a consequence, the effectiveness of item to item approaches may vary and more investigations are required to address this issue and assess the generality over a greater number of domains and testing conditions.

Returning now to the qualities that we look for in a student modeling framework and that we outlined in the introduction, and notwithstanding the issues we discussed, we conclude that item to item structures offer a great potential. The experiments we conducted showed their effectiveness for performing knowledge assessment with models learned from very small data sets (as few as 30 data cases for the Unix experiment with POKS). Yet, they display all the advantages of learned graph probabilistic models, namely the efficient automation of building fine grained model and the waiving of human intervention, which forgo the human expertise bottleneck and subjectivity bias, and offers the possibility of estimating the reliability of the diagnostic. The POKS framework also has the quality of being grounded in the theory of knowledge spaces and in the mathematically simple naive Bayes frame-

work. Finally, the technique of augmenting an initial set of evidence using an item to item structure is a generic means to complement other models of skill and concept assessment. It fits into the perspective of reusing generic tools for student modeling.

Acknowledgements

We are endowed to the anonymous reviewers and to Richard Labib for their valuable feedback on earlier drafts of the paper. We are also grateful to Jiří Vomlel for providing the data used in this experiment This work has been supported by the National Research Council of Canada.

References

Aleven, V., B. M. McLaren, J. Sewall, and K. R. Koedinger: 2006, 'The cognitive tutor authoring tools (CTAT): versatile and increasingly rapid creation of tutors'. In: *ITS'06: Proceedings of the 8th International Conference on Intelligent Tutoring Systems*.

Carmona, C., E. Millán, J.-L. P. de-la Cruz, M. Trella, and R. Conejo: 2005, 'Introducing Prerequisite Relations in a Multi-layered Bayesian Student Model.'. In: L. Ardissono, P. Brna, and A. Mitrovic (eds.): *User Modeling*, Vol. 3538 of *Lecture Notes in Computer Science*. pp. 347–356.

Chow, C. and C. Liu: 1968, 'Approximating discrete probability distributions with dependence trees'. *IEEE Trans. Information Theory* **14**(11), 462–467.

Conati, C., A. Gertner, and K. VanLehn: 2002, 'Using Bayesian Networks to Manage Uncertainty in Student Modeling'. *User Modeling and User-Adapted Interaction* **12**(4), 371–417.

Cooper, G. F. and E. Herskovits: 1992, 'A Bayesian Method for the Induction of Probabilistic Networks from Data'. *Machine Learning* **9**, 309–347.

Desmarais, M. C., A. Maluf, and J. Liu: 1996, 'User-Expertise Modeling with Empirically Derived Probabilistic Implication Networks'. *User Modeling and User-Adapted Interaction* **5**(3-4), 283–315.

Desmarais, M. C. and X. Pu: 2005, 'A Bayesian Inference Adaptive Testing Framework and its comparison with Item Response Theory'. *International Journal of Artificial Intelligence in Education* **15**, 291–323.

Doignon, J.-P. and J.-C. Falmagne: 1999, *Knowledge Spaces*. Berlin: Springer-Verlag.

Domingos, P. and M. Pazzani: 1997, 'On the optimality of the simple Bayesian classifier under zero-one loss'. *Machine Learning* **29**, 103–130.

Dösinger, G.: 2002, 'Adaptive Competence Testing in eLearning'. *European Journal of Open and Distance Learning, http://www.eurodl.org/* p. 12 (online publication).

Dowling, C. E. and C. Hockemeyer: 2001, 'Automata for the Assessment of Knowledge'. *IEEE Transactions on Knowledge and Data Engineering* **13**(3), 451–461.

Falmagne, J.-C., E. Cosyn, J.-P. Doignon, and N. Thiéry: 2006, 'The Assessment of Knowledge, in Theory and in Practice'. In: R. Missaoui and J. Schmid (eds.): *ICFCA*, Vol. 3874 of *Lecture Notes in Computer Science*. pp. 61–79.

Falmagne, J.-C., M. Koppen, M. Villano, J.-P. Doignon, and L. Johannesen: 1990, 'Introduction to knowledge spaces: How to build test and search them'. *Psychological Review* **97**, 201–224.

François, O. and P. Leray: 2003, 'Etude comparative d'algorithmes d'apprentissage de structure dans les réseaux bayésiens.'. In: *Proceedings of RJCIA03, plateforme AFIA03*. pp. 167–180.

Friedman, N., D. Geiger, and M. Goldszmidt: 1997, 'Bayesian Network Classifiers'. *Mach. Learn.* **29**(2-3), 131–163.

Heller, J., C. Steiner, C. Hockemeyer, and D. Albert: 2006, 'Competence–Based Knowledge Structures for Personalised Learning'. *International Journal on E–Learning* **5**(1), 75–88.

Horvitz, E.: 1999, 'Principles of Mixed-Initiative User Interfaces'. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI'09*. pp. 159–166.

Jensen, F. V.: 1996, *An introduction to Bayesian Networks*. London, England: UCL Press.

Kambouri, M., M. Koppen, M. Villano, and J.-C. Falmagne: 1994, 'Knowledge assessment: tapping human expertise by the QUERY routine'. *International Journal of Human-Computer Studies* **40**(1), 119–151.

Kay, J., B. Kummerfeld, and P. Lauder: 2002, 'Personis: A Server for User Models'. In: P. D. Bra, P. Brusilovsky, and R. Conejo (eds.): *AH*, Vol. 2347 of *Lecture Notes in Computer Science*. pp. 203–212.

Kobsa, A.: 2001, 'Generic User Modeling Systems'. *User Model. User-Adapt. Interact* **11**(1-2), 49–63.

Kodaganallur, V., R. R. Weitz, and D. Rosenthal: 2005, 'A Comparison of Model-Tracing and Constraint-Based Intelligent Tutoring Paradigms'. *International Journal of Artificial Intelligence in Education* **15**, 117–144.

Martin, B. and A. Mitrovic: 2001, 'Easing the ITS Bottleneck with Constraint-Based Modelling'. *New Zeleand Journal of Computing* **8**(3), 38–47.

Millán, E., M. Trella, J.-L. Pérez-de-la-Cruz, and R. Conejo: 2000, 'Using Bayesian Networks in Computerized Adaptive Tests'. In: M. Ortega and J. Bravo (eds.): *Computers and Education in the 21st Century*. Kluwer, pp. 217–228.

Mitrovic, A., M. Mayo, P. Suraweera, and B. Martin: 2001, 'Constraint-Based Tutors: A Success Story'. In: L. Monostori, J. Váncza, and M. Ali (eds.): *IEA/AIE*, Vol. 2070 of *Lecture Notes in Computer Science*. pp. 931–940.

Murphy, K. P.: 2001, 'The Bayes Net Toolbox for MATLAB'. Technical report, University of California at Berkeley; Berkeley , CA.

Neapolitan, R. E.: 2004, *Learning Bayesian Networks*. New Jersey: Prentice Hall.

Pearl, J.: 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. and T. Verma: 1991, 'A Theory of Inferred Causation'. In: *Knowledge Representation and Reasoning: Proceedings of the Second International Conference*. pp. 441–452.

Rish, I.: 2001, 'An Empirical Study of the Naive Bayes Classifier'. In: *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*. pp. 41–46.

Spirtes, P., C. Glymour, and R. Scheines: 2000, *Causation, Prediction, and Search*. Cambridge, Massachusetts: The MIT Press, 2 edition.

Suraweera, P., A. Mitrovic, and B. Martin: 2005, 'A knowledge acquisition system for constraint-based intelligent tutoring systems'. In: C. Looi, G. McCalla, B. Bredeweg, and J. Breuker (eds.): *Proceedings of the 12th International Conference on Artificial Intelligence in Education, AEID'2005*. Amsterdam, pp. 638–645.

Tatsuoka, K.: 1983, 'Rule space: An approach for dealing with misconceptions based on item response theory'. *Journal of Educational Measurement* **20**, 345–354.

VanLehn, K., C. Lynch, K. Schulze, J. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill: 2005, 'The Andes Physics Tutoring System: Five Years of Evaluation'. In: *Proceedings of AIED'05*. pp. 678–685.

VanLehn, K., Z. Niu, S. Siler, and A. S. Gertner: 1998, 'Student Modeling from Conventional Test Data: A Bayesian Approach Without Priors'. In: *ITS'98: Proceedings of the 4th International Conference on Intelligent Tutoring Systems*. London, UK, pp. 434–443.

Vomlel, J.: 2004, 'Bayesian networks in educational testing'. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* **12**(Supplementary Issue 1), 83–100.

Wang, W.-C. and P.-H. Chen: 2004, 'Implementation and Measurement Efficiency of Multidimensional Computerized Adaptive Testing'. *Applied Psychological Measurement* **28**(5), 295–316.

Zapata-Rivera, J.-D. and J. Greer: 2004, 'Inspectable Bayesian student modelling servers in multi-agent tutoring systems'. *Int. J. Hum.-Comput. Stud.* **61**(4), 535–563.

Appendix A

Proof that A → B → C implies that there is at least one symmetric relation in this cycle

This appendix addresses the question of whether POKS can have a cycle other than through symmetric relations, such as the following cyclic structure:

$$A \rightarrow B \rightarrow C \rightarrow A$$

We show that such cycle cannot be found without a symmetric relation.

First, we can readily assume that the interaction test (inequality 3) is positive for all three pairs A-B, B-C, and A-C of this structure.

Next, assume the following contingency table :

		A T	**A** F
B	T	ab_1	ab_2
	F	ab_3	ab_4

where ab_x refers to a frequency count for the co-occurrence of A and B corresponding to their respective presence (T) and absence (F).

Let us suppose that there is no symmetric relation in this cycle, that is, none of the edges $B \rightarrow A$, $C \rightarrow B$ and $A \rightarrow C$ exists. Since $A \rightarrow B$, and we do not have $B \rightarrow A$, then

$$P(B|A) = ab_1/(ab_1 + ab_3) \geq p_c$$
$$P(A|B) = ab_1/(ab_1 + ab_2) < p_c$$
$$ab_1/(ab_1 + ab_3) > ab_1/(ab_1 + ab_2)$$

From this we may conclude that $ab_3 < ab_2$ and $frequency(B) > frequency(A)$. Similarly, we can conclude $frequency(C) > frequency(B)$ and, by transitivity, $frequency(C) > frequency(A)$.

Let us now consider the contingency table for variables A and C:

		A	
		T	F
C	T	ac_1	ac_2
	F	ac_3	ac_4

The relation $C \rightarrow A$ implies that $P(A|C) = ac_1/(ac_1 + ac_2) \geq p_c$. Since $frequency(C) > frequency(A)$, we know that $ac_1 + ac_2 > ac_1 + ac_3$. Thus, we obtain $P(C|A) = ac_1/(ac_1 + ac_3) > ac_1/(ac_1 + ac_2) \geq p_c$. This result implies that the first condition (inequality 1) for the relation $A \rightarrow B$ is satisfied.

A similar demonstration can be done to show that the second condition can also be satisfied (inequality 2 corresponding to $P(\overline{A}|\overline{B})$). Given that the third condition (the interaction test) can readily be assumed satisfied, we can conclude that there must exist a relation $A \rightarrow C$, which contradicts our assumption. Therefore, there must be at least one symmetric relation in the cycle.

Authors' Vitae

Michel C. Desmarais
is Assistant Professor at the Computer Engineering Department of Polytechnique Montreal since 2002. He received his Ph.D. degree in psychology in 1990 from the University of Montreal. He was team leader of the HCI and Learning Environments groups at the Computer Research Institute of Montreal between 1990 and 1998, where he was involved in a number of research projects in close collaboration with private corporations. From 1998 to 2002, he directed a number of software development projects in a private company. His research interests are in user modeling, e-learning, human-computer interactions, and soft-

ware engineering. He authored of over 50 papers in scientific journal, conferences, and book chapters.

Michel Gagnon
is Assistant Professor at the Computer Engineering Department of Polytechnique Montreal since 2002. He received his Ph.D. degree in computer science in 1993 from the University of Montreal. Since then, he has been working on natural language engineering, with a special attention to semantics. From 1995 to 1998, he participated in many projects at Machina Sapiens inc., a company which at that time was a leader in the development of grammar checkers. Since 2002, his reasearch activities also include the semantic web, with special attention to e-learning applications.

Peyman Meshkinfam
is a Ph.D. candidate in Ecole Polytechnique De Montreal interested in working on automated test data generation and tools development as well as reverse engineering in software engineering. He received his master degree in computer engineering from Ecole Polytechnique de Montreal in 2005. During his research he worked on comparison of different probabilistic network approaches in the domain of knowledge assessment. He received his master degree in Telecommunication from Ecole Polytechnique de Montreal in 1998 and his bachelor in solid-state physics from Tehran University in 1988. He has 8 years of industry experience in reliability quality and testing.